

Genome scans for detecting footprints of local adaptation using a Bayesian factor model

Nicolas Duforet-Frebourg¹, Eric Bazin², Michael G.B. Blum¹

¹ Université Joseph Fourier, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG, UMR 5525, Grenoble, France.

² Université Joseph Fourier, Centre National de la Recherche Scientifique, Laboratoire d'Ecologie Alpine, UMR 5553, Grenoble, France.

Running Head: Genome scans based on Bayesian factor model

Keywords: F_{ST} , population structure, landscape genetics, population genomics, selection scan

Corresponding author: Michael Blum

Laboratoire TIMC-IMAG, Faculté de Médecine, 38706 La Tronche, France

Phone +33 4 56 52 00 65

Fax +33 4 56 52 00 55

Email: michael.blum@imag.fr

Abstract

A central part of population genomics consists of finding genomic regions implicated in local adaptation. Population genomic analyses are based on genotyping numerous molecular markers and looking for outlier loci in terms of patterns of genetic differentiation. One of the most common approach for selection scan is based on statistics that measure population differentiation such as F_{ST} . However they are important caveats with approaches related to F_{ST} because they require grouping individuals into populations and they additionally assume a particular model of population structure. Here we implement a more flexible individual-based approach based on Bayesian factor models. Factor models capture population structure with latent variables called factors, which can describe clustering of individuals into populations or isolation-by-distance patterns. Using hierarchical Bayesian modeling, we both infer population structure and identify outlier loci that are candidates for local adaptation. As outlier loci, the hierarchical factor model searches for loci that are atypically related to population structure as measured by the latent factors. In a model of population divergence, we show that the factor model can achieve a 2-fold or more reduction of false discovery rate compared to the software BayeScan or compared to a F_{ST} approach. We analyze the data of the Human Genome Diversity Panel to provide an example of how factor models can be used to detect local adaptation with a large number of SNPs. The Bayesian factor model is implemented in the open-source PCAdapt software.

Introduction

With the development of sequencing and genotyping technologies, there is a considerable impetus to pinpoint loci involved in local adaptation (Akey et al. 2002; Bonin et al. 2006). Finding genomic regions subject to local adaptation is a central part of “population genomics”, which is based on genotyping numerous molecular markers and looking for outlier loci (Luikart et al. 2003). The main principle of population genomics is that most loci have neutral patterns of variation that are similarly affected by demographic processes whereas loci targeted by natural selection have atypical patterns. Measures of genetic differentiation between populations such as F_{ST} have been commonly used to find outlier loci although there are many alternative approaches (Oleksyk et al. 2010). Loci involved in local adaptation have increased values of genetic differentiation for populations living in different environments. A proof of concept was provided when studying human adaptation to altitude because the most differentiated variants between a Tibetan population living in a hypoxic environment and a lowland Han Chinese population were found in hypoxia-inducible transcription factors (Yi et al. 2010; Xu et al. 2011).

Genome scans based on F_{ST} were proposed by Lewontin and Krakauer (1973) and have been considerably expanded since (Beaumont and Nichols 1996; Vitalis et al. 2001; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler et al. 2008; Guo et al. 2009; Bonhomme et al. 2010; Bazin et al. 2010; Gompert and Buerkle 2011; Fariello et al. 2013). They are not limited to two populations as in the adaptation-to-altitude example and can be used with multiple populations. One possibility is to compute an overall F_{ST} measure of genetic differentiation and to determine a threshold at which the null hypothesis of neutral evolution can be rejected (Beaumont and Nichols 1996). Another possibility is to adopt a Bayesian perspective by implementing the multinomial-Dirichlet model or F-model, which is parametrized by population-specific F -statistics (Beaumont and Balding 2004). The F -statistics can be interpreted as measures of divergence from a common immigrant gene pool (Wright 1931) or as divergence from an initial and hypothetical ancestral population (Nicholson et al. 2002). The Bayesian approach for distinguishing between neutral or adaptive evolution offers the opportunity to assign a probability to each of the two evolutionary models at each locus (Foll and Gaggiotti 2008; Riebler et al. 2008). In the following, we refer to genome scans based on F -statistics for approaches based on F_{ST} or based on population-specific F -statistics. There are many software implementing genome scans based on F -statistics (e.g. *BayeScan*, *DetSel*, *fdist2*, *Lositan*), and they contribute to the popularity of this approach in population genomics (Beaumont and Nichols 1996; Vitalis et al. 2003; Antao et al. 2008; Foll and Gaggiotti 2008).

However, a major issue with genome scans based on F -statistics is that they can generate a high rate of false-positives for both biological and statistical reasons (Bierne et al. 2013). Here we propose to address the statistical and computational problems that arise with F -statistics. The first problem arises because F -statistics have been derived under the Wright’s F -model of population subdivision, which assumes a particular covariance structure for

gene frequencies among populations (Bierne et al. 2013; Fourcade et al. 2013). When spatial structure departs from Wright’s island model of population subdivision, genome scans based on F -statistics produce many false positives and alternative statistical measures that account for population structure have recently been proposed (Bonhomme et al. 2010; Günther and Coop 2013). A second potential problem concerns the computational burden of some Bayesian approaches, which can become an obstacle with large number of SNPs (Lange et al. 2014). The last intrinsic problem of genome scans based on F -statistics is that individuals should be grouped into populations. However, it has been advocated in landscape genetics to rather work at the scale of individuals because it avoids potential bias in identifying populations in advance and it offers the opportunity to conduct studies at a finer scale (Manel and Holderegger 2013).

To tackle the aforementioned problems, we propose a statistical method based on a Bayesian factor model (West 2003) to pick outlier loci involved in local adaptation. With factor models, we seek to jointly determine population structure and outlier loci. Factor models are strongly related to principal components analysis (PCA) because they both approximate the matrix of individual genotypes by a product of two lower-rank matrices, albeit using different constraints and priors for the lower-rank matrices (Engelhardt and Stephens 2010). One of the two matrices encodes population structure using latent factors whereas the second matrix measures to what extent each individual SNP is related to the pattern of population structure. The proposed factor model seeks for loci that are atypically related with population structure. To show the potential of factor models for genome scan, we consider two examples. First, we consider a model of population divergence. In this example, we compare false discovery rates obtained with the proposed factor model, with *BayeScan*, and with a genome-scan based on F_{ST} . The second example is a model of isolation-by-distance with selection. It is an instance of how factor models can be used to detect local adaptation when it would be arbitrary to group individuals into populations. Last, we analyze the HGDP human data-set (Li et al. 2008) to provide an example of how factor models can be used to detect local adaptation with a large number of SNPs.

New Approaches

We denote by \mathbf{G} the $n \times p$ matrix of allele counts where n is the number of individuals and p is the number of loci. The elements $G_{i\ell}$, $i = 1, \dots, n$, $\ell = 1, \dots, p$, correspond to the allele counts of the i^{th} individual at locus ℓ and belong to $\{0, 1\}$ or $\{0, 1, 2\}$ for haploid and diploid species respectively. We assume that the matrix of genotypes has been centered (each column has a mean of 0), and the resulting matrix is denoted by \mathbf{Y} .

Factor models assume that the matrix of column-centered genotypes \mathbf{Y} can be written as a product of two lower-rank matrices \mathbf{U} and \mathbf{V} of dimension $(n \times K)$ and $(K \times p)$ where K is an hyper parameter of the factor model. Denoting by $\mathbf{U}_1, \dots, \mathbf{U}_K$ the column-vectors of \mathbf{U} referred as *factors* or *latent factors* in the following, factor models

assume that the vector—of size n —of centered allele counts \mathbf{Y}_ℓ can be obtained as

$$\mathbf{Y}_\ell = \sum_{k=1}^K \mathbf{U}_k V_{k\ell} + \boldsymbol{\epsilon}_\ell, \ell = 1 \dots, p, \quad (1)$$

where $\boldsymbol{\epsilon}_\ell$ is a vector containing n independent Gaussian residuals of variance σ^2 and where the $V_{k\ell}$ are the elements of the matrix \mathbf{V} . Assuming that the K factors are known, then the elements $V_{k\ell}$ of the matrix \mathbf{V} are the regression coefficients obtained after regressing the vector of centered allele counts \mathbf{Y}_ℓ by the K factors $\mathbf{U}_1, \dots, \mathbf{U}_K$. The outlier approach we advocate is to consider as candidates for local adaptation the loci ℓ that have large (in absolute value) regression coefficients $V_{k\ell}$ for one of the factor $\mathbf{U}_1, \dots, \mathbf{U}_K$. In factor models, the K factors $\mathbf{U}_1, \dots, \mathbf{U}_K$ are in fact unknown and have to be estimated; they are parameters of the model and represent population structure (Engelhardt and Stephens 2010). In our proposed framework, outlier loci are therefore loci that are excessively related with population structure, as measured by the latent factors. After statistical inference, the factors $\mathbf{U}_1, \dots, \mathbf{U}_K$ are ordered by decreasing variances $\sigma_1^2 > \dots > \sigma_K^2$ where σ_k^2 measures the variance of the regression coefficients $V_{k\ell}$ for the k^{th} factor.

To provide a concrete example of how factors represent population structure, we consider a model of population divergence. We assume that an initial population splits into two populations A and B that diverged according to neutral evolution. The initial neutral divergence of duration T is followed by 2 concomitant splits where each daughter population A and B splits into two subpopulations (A_1, A_2) and (B_1, B_2) . By contrast to the initial divergence that is purely neutral, the second phase of divergence between populations assumes some local adaptation with a small proportion of SNPs conferring selective advantage (Figure 1). We fit the factor models with $K = 3$ and we display the three factors in Figure 1. The first factor discriminates individuals according to the initial split and the second and third factor discriminate individuals according to the subsequent splits which separate subpopulation A_1 from A_2 (second factor) and subpopulation B_1 from B_2 (third factor).

We now specify how we measure in a Bayesian fashion the degree of outlyingness for each locus. To account for outlier and non-outlier loci, we assume that, at a given locus ℓ , the vector of regression coefficients $\mathbf{V}_\ell = (V_{1\ell}, \dots, V_{K\ell})$ comes from a mixture of two different distributions. We introduce a vector \mathbf{z} of indicator variables (z_1, \dots, z_p) whose elements are equal to 0 for non-outlier loci and take values in $1, \dots, K$ for outlier loci. For both non outlier and outlier locus ℓ , we assume that the vector $\mathbf{V}_\ell = (V_{1\ell}, \dots, V_{K\ell})$ is composed of independent Gaussian random variables. The model for non-outlier loci is a product of Gaussian distributions

$$V_{k\ell} | z_\ell = 0 \sim \mathcal{N}(0, \sigma_k^2), \quad k = 1, \dots, K, \quad (2)$$

where $\mathcal{N}(m, \sigma^2)$ denotes the Gaussian distribution of mean m and variance σ^2 . To model outlier loci, we consider

a variance-inflation model which assumes an inflated variance to account for outlier loci (Box and Tiao 1968; Devlin and Roeder 1999). The model for outlier loci is itself a mixture model with K components of equal weights where the k^{th} component assumes an inflated variance for the k^{th} regression coefficient but not for the other ones. Denoting by c_k^2 the variance-inflation parameter for factor k ($c_k^2 > 1$), the k_0^{th} component of the mixture model for outlier assumes a product of Gaussian distributions

$$\begin{aligned} V_{k\ell}|z_\ell = k &\sim \mathcal{N}(0, c_k^2 \sigma_k^2), \quad k = k_0 \\ V_{k\ell}|z_\ell = k &\sim \mathcal{N}(0, \sigma_k^2), \quad k \neq k_0 \end{aligned} \quad (3)$$

The model for outliers has been chosen for sake of interpretability. Each outlier locus can be related to one of the K factors because outlier loci should be atypically explained by one of the K factors. To measure for each locus the strength of evidence for outlyingness, we compute the Bayes factor of the outlier model against the non-outlier model. If a locus is considered as an outlier, the factor with which there is an atypical correlation is found by computing the posterior probabilities of each of the K components of the outlier mixture model. To account for Linkage Disequilibrium (LD), we additionally consider a Potts Model that encourages outlier loci to be clustered in the genome (see METHODS for details).

When fitting the factor model with $K = 3$ to data simulated under the scenario of population divergence depicted in Figure 1, the outlier model of equation (3) assumes 3 different types of outlier loci : loci that have large genetic differentiation when comparing the pair of subpopulations (A_1, A_2) to the pair (B_1, B_2) (large values of $|V_{1\ell}|$), loci that have large genetic differentiation when comparing subpopulation A_1 to A_2 (large values of $|V_{2\ell}|$), and loci that have large genetic differentiation when comparing subpopulation B_1 to B_2 (large values of $|V_{3\ell}|$). Because the simulation assumes that the initial period of divergence is purely neutral, the first types of outliers (large values of $|V_{1\ell}|$) are in fact false positives.

Results

Simulation study

Population divergence model

The first simulation study investigates to what extent factor models better account for population structure than methods based on F -statistics. We consider the model of population divergence depicted in Figure 1. An initial neutral divergence is followed by adaptive divergence where 4% of the 10,000 simulated SNPs are involved in local adaption. The set of adaptive SNPs is split in four equal parts and each subset of SNP confers a selective advantage in only one

of the four populations. When the initial neutral divergence time T is null, the population tree is star-like, and the assumption of the F -model is valid. As the initial neutral divergence time T increases, the departure to the F -model increases. The neutral divergence time T is scaled so that $T = 1$ means that the neutral and adaptive phases are of same duration.

First, we present results using the factor model with $K = 3$ factors that is optimal because there are 4 populations in the divergence model (Patterson et al. 2006). We consider a long-enough divergence time $T = 2$ so that the first factor corresponds to the initial and neutral divergence whereas the second and third factors correspond to the subsequent divergence events during which biological adaptation took place (Figure 1). The SNPs that have been truly involved in biological adaptation are usually associated with the correct factor because among the 400 truly adaptive SNPs, 81% are associated with the second and third factor and this proportion raises to 98% (resp. 92%) when considering the 195 (resp. 305) adaptive SNPs with Bayes factors larger than 10 (resp. 1) (Figure 2).

Then, we compare the false discovery rates of three different approaches including the proposed factor model, *BayeScan* (version 2.1), and genome scans based on the F_{ST} statistic. For both *BayeScan* and the proposed factor model, we use Bayes factors for ranking SNPs whereas we use F_{ST} values for the last method. More precisely, we use the q-values for ranking SNPs with *BayeScan* but by definition of the q-values, it provides the same ranking as the Bayes factors. To determine a threshold above which a SNP is considered as an outlier, we enlarge the lists of top ranked SNPs, provided by each method, until each of them contains 50% of the 400 truly adaptive SNPs. This procedure amounts at setting the sensitivity to 50% (the sensitivity is also called recall rate in machine learning). Figure 3 shows that for all methods, the false discovery rate (FDR) is below 5% when the population tree is almost star-like ($T = 0.04$) but it increases with the initial neutral divergence time T . Although the FDR always increases with T , the FDRs of the factor model are always smaller than FDRs obtained with F_{ST} and with *BayeScan*. For instance, when $T = 1$, the FDRs obtained with F_{ST} and *BayeScan* are between 20% and 30% whereas it is smaller than 5% with the factor model. Instead of using a threshold of 50%, we also constrain the lists of SNPs to contain 25% or 75% of the truly adaptive SNP (i.e. setting the sensitivity to 25% or 75%). As for the 50% threshold, all methods have small FDR for small-enough initial divergence time T , and, as T increases, FDR increases at a slower rate for the factor model (Figure S1). In summary, the false discovery rate increases as the model of divergence deviates from a star-like phylogeny, but compared to other methods, the factor model reduces the proportion of false discoveries by a factor of 2 or more when there is a strong-enough deviation from the star-like assumption ($T > 0.8$).

The results presented so far were obtained using the factor model with $K = 3$ factors. By increasing the values of K from 1 to 6, we find that, compared to $K = 3$, the false discovery rate drastically increases for underspecification of K ($K < 3$) but is almost insensitive to overspecification of K ($K > 3$, Figure S2). We also compute the mean square error (MSE) of equation (1) for different values of K to determine if the MSE can be a guide for choosing K . The MSE decreases from $K = 1$ to $K = 3$ before staying almost constant as K continues to grow (Figure S3). In this

example of population divergence, the MSE suggests to choose $K = 3$ but choosing a more complex model ($K > 3$) would provide comparable false discovery rates.

Isolation-by-distance model

The second simulation study provides an example of how to search for biological adaptation when there is isolation-by-distance. Approaches based on F -statistics would require to group individuals into populations, and we want to avoid that. On a two-dimensional 10×10 grid, we simulate a stepping-stone model with selection acting on individuals located in the lower-right corner of the grid. We sample 10 diploid individuals at each of the 100 demes. A total of 50 out of 2,050 SNPs confer selective advantage in this region and the selection coefficient decreases gradually when moving away from the point where selection is maximal.

With the factor model, the selection gradient is reflected in a different factor depending on the value of the selection coefficients (results not shown). Here we choose the intensity of selection such as the selection gradient is visible in the third factor (Figure 4). The other factors have spatial patterns that are typical of isolation-by-distance models (Novembre and Stephens 2008). We choose $K = 4$ because the MSE decreases from $K = 1$ to $K = 4$ before being almost constant (Figure S3). In terms of false discovery rate, this choice of K is not optimal because $K = 3$ would provide smaller FDR (Figure S4). However, as in the first example, smaller values of K compared to the optimal value ($K < 3$) increases FDR drastically whereas too large values of K ($K > 3$) increases the optimal FDR more moderately (Figure S4). With $K = 4$, the FDR is of 0% when considering the top 25 SNPs, which corresponds to a sensitivity of 50%. However, when setting the sensitivity at 75%, the FDR increases to 30%, which corresponds to 38 true positive SNPs among a list of 54 SNPs. The 50 truly adaptive SNPs are all correctly associated with the factor corresponding to biological adaptation, which is the third factor here (Figure 2). When decreasing the number of sampled individuals from 10 to 1, the false discovery rate, obtained with a sensitivity of 50%, increases considerably from 0% to 91% (Figure S5).

Analysis of human SNP data

The HGDP dataset contains 644,199 SNPs, after removal of the SNPs on the sex chromosomes and on the mitochondrion, which have been typed for 1,043 individuals coming from 53 different populations (Li et al. 2008). First, we fit the factor model with different values of K . By contrast with the two previous examples, there is no value of K at which the MSE stops to decrease (Figure S3). By looking at the different factors (Figure 5 and Figure S6), we decide to consider a model that captures genetic differentiation between but not within continents. Using this criterion, we consider a factor model with $K = 4$ since larger values of K would reveal genetic difference within continents (Figure S6). The first factor mostly contrasts African from Asiatic and Native American individuals and the second factor

mainly discriminates African from Middle-Eastern and Western Eurasian individuals. The third factor distinguishes Native American individuals—coming from Central and South America—from the rest of the sample whereas the last factor separates individuals from Oceania from the rest of the sample (Figure 5).

We choose to restrict our analysis to the 5,000 top-hit SNPs (Table S1). Their values of the Bayes factors range from 1.03 to 5.05 on a \log_{10} scale. The two SNPs with the largest Bayes factors (rs1834640 and rs2250072) are correlated with the second factor. They are located on chromosome 15 and the closest gene is *SLC24A5*, which is located at 20 – 30 kb from the SNPs. Among the 5,000 SNPs with largest Bayes factors, 851 are related to factor 1, 844 with factor 2, 1,982 with factor 3 and 1,323 with factor 4. For each of the four sublists, we further provide information for the 10 SNPs with the largest Bayes factors (Table 1).

- For the first factor, although we consider ten different SNPs, only two genomic regions are found. One of the two genomic regions is located on chromosome 10 and downstream of the oncogene *CYP26A1* whose expression is enhanced in sunlight-damaged human skin (Osanai and Lee 2011; Mallick et al. 2013). The other SNPs were found in the *SM6* gene which is implicated in the structural maintenance of chromosome protein 6 and which has already been picked as a candidate for selection in another scan with the HGDP sample (Hao et al. 2013). For all the ten SNPs, we investigate the worldwide repartition of allele frequencies with the ALFRED database (Rajeevan et al. 2012). East Asiatic and Native American populations have allele frequencies that are different from the rest of the sample (Table S2) as can be predicted when looking at the geographic repartition of the first factor (Figure 5).
- For the outlier SNPs associated with the second factor, the allelic frequencies were mostly different when comparing Western Eurasian individuals to the rest of the sample (Table S2). In addition to the SNPs close to the *SLC24A5* gene that is associated with light skin in Western Eurasia (Canfield et al. 2013), we also find four other regions locating close to the following genes: *EDAR* in chromosome 2 which has been associated with various traits including hair thickness and sweating (Kamberov et al. 2013), *SLC35F3* in chromosome 1, *KIF3A* in chromosome 5, *RABGAP1* and *STRBP* in chromosome 9 with the latter being involved in spermatogenesis, and *MYO5C* and *DUT* in chromosome 15.
- For the third factor, eight out of the ten SNPs with the largest Bayes factor are found in a 1Mb region of chromosome 22, which encompasses many different genes (Table 1). For the SNPs in this large region of chromosome 22, the allelic frequencies mostly differ between Native Americans and the rest of the sample. For Sub-Saharan African populations, allelic frequencies of these SNPs are intermediate with Pygmies populations having frequencies that are often the most similar to the Native Americans (Table S2).
- The allele frequencies of the SNPs that are the most associated with the fourth factor mostly differ between

individuals from Oceania (Papuan and Melanesian) and the rest of the sample with Native Americans and Pygmies population having, for some SNPs, allele frequencies that are the most similar to the Oceanians (Table S2). Among the ten outlier SNPs, four SNPs are located in chromosome 8 and four SNPs are located in chromosome 17. Among the 1,323 SNPs associated with the fourth factor, there is an excess of outlier SNPs in chromosome 8 (Figure S7) pointing to a prominent role of its genes in adaptation to the local conditions of Oceania. There are different genomic regions with large Bayes factors in chromosome 8 and one of these genomic regions encompasses *RPILI*, a gene often found in selection scan (Barreiro et al. 2008) and related to eye diseases (Davidson et al. 2013).

We also perform a Gene Ontology (GO) enrichment analysis on human genes using the 5,000 SNPs with the largest Bayes factors. We find significant enrichment of biological processes for each of the four factors (Table S3). Some interesting instances of the enriched gene ontologies include three different GO terms related to regulation of hormone secretion for the first factor, enrichment of homophilic cell adhesion for the third factor, and aging for the fourth factor. Finally, we look at a catalog of published GWAS (Welter et al. 2014) to search for enrichment of outlier SNPs related to a particular phenotype (Table S4). The traits that are the most associated with the outlier SNPs are height (6 SNPs), obesity and weight (5 SNPs) and Crohn’s disease (5 SNPs).

Discussion

Based on a Bayesian factor model, we provide a new approach for performing genome scan of local adaption. The hierarchical factor model considers as outliers the SNPs that are atypically related to population structure. A set of K latent factors measure population structure and they can adequately describe clustering of individuals into populations (Figure 1), isolation-by-distance patterns and gradients of selection (Figure 4). Compared to the software *BayeScan* or to genome scans based on F_{ST} , the factor model does not assume a particular model of population structure. In a model of population divergence, we show that removing the assumptions of the F -model considerably reduces the false discovery rate. To explain why the factor model generates less false discoveries, we introduce the notions of mechanistic and phenomenological models (Hilborn and Mangel 1997). Mechanistic models aim to mimic the biological processes that are thought to have given rise to the data whereas phenomenological models seek only to best describe the data using a statistical model. In the spectrum between mechanistic and phenomenological model, the F -model would stand close to mechanistic models whereas factor models would be closer to the phenomenological ones. Mechanistic models are appealing because they provide quantitative measures that can be related to biologically meaningful parameters. For instance, F -statistics measure genetic drift which can be related to migration rates, divergence times or population sizes. By contrast, phenomenological models work with mathematical abstractions such as latent factors that can be difficult to interpret biologically. The downside of mechanistic models is that violation of

the modeling assumption can invalidate the proposed framework and generate many false discoveries in the context of selection scan. The F -model assumes a particular covariance matrix between populations which is found with star-like population trees for instance (Bonhomme et al. 2010). However, more complex models of population structure can arise for various reasons including non-instantaneous divergence and isolation-by-distance, and they will violate the mechanistic assumptions and make phenomenological models preferable.

Although principal component analysis or the related factor model are generally used to investigate population structure, there have been already several attempts at performing selection scan based on these statistical approaches. A first idea is to compute F_{ST} values between pairs of populations that contain the top and bottom individuals for each principal component (Abdellaoui et al. 2013). This approach provides a list of outliers that are specific to each principal component in the same way as the hierarchical factor model of equations (1)-(3) provides outliers that are related to one of the K factors. A second proposition involves new interpretations of PCA related to F -statistics, which provide statistical measures to detect local adaptation (Laloë and Gautier 2011). A last and recent proposition called “logistic factor analysis” adds a logistic link function to the factor model (equation (1)) in order to guarantee that the predicted values can be interpreted as frequencies because they lie between 0 and 1 (Hao et al. 2013). Loci involved in biological adaptation were scanned using a deviance statistic (Hao et al. 2013). These related approaches are built on the success of PCA and factor models to capture population structure with a small number of variables.

Choosing the dimension K of the statistical model that ascertains population structure is a recurrent problem. One possibility is to use an *objective* approach based on a quantitative criterion. Examples of such objective criteria include the Δ_K measure to detect the number of clusters using the software STRUCTURE (Evanno et al. 2005) or the Tracy-Widom statistic to choose the dimension K in PCA (Patterson et al. 2006). Another possibility is to adopt a *subjective* approach and to choose a value of K such that increasing K would provide results that are considered of too little interest. With the proposed Bayesian factor model, we implemented both approaches. For the simulations, choosing K based on the mean square error (MSE) of equation (1) works well because the MSE stops to decrease when K increases beyond a certain value. However, for the human data, the choice is more complex because the MSE decreases regularly as K increases. We chose $K = 4$ because we were only interested in biological adaptation that is related to genetic differentiation between continents but we acknowledge that major adaptive processes although occurs within continents (Jarvis et al. 2012). To provide recommendations for choosing K , we suggest to fit the hierarchical factor model with different values of K in order to investigate if there is a value of K , at which the MSE stops or almost stops to decrease. If not, the choice of K can be based on subjective arguments where the latent factors of too little interest can be discarded.

One of our objective was to propose a method for selection scan that avoids the computational burden of some Bayesian approaches, which can become a serious obstacle when analyzing large scale SNP data. This objective is however only partly fulfilled. The downside of our approach is that it relies on a MCMC algorithm that quickly grinds

to a halt under the sheer mass of SNP data (Lange et al. 2014). Hopefully, the MCMC algorithm is based on a Gibbs sampler that alternates the computation of least square solutions, which are fast to compute. For the HGDP dataset (644,199 SNPs), the run-time ranges from 13 to 16 hours using a single computer processor (2.4 GHz 64bit Intel Xeon) when K increases from 1 to 8. Because dataset containing millions of SNPs are becoming available, we are currently working on the development of a faster version of our software.

Fitting the factor model to the HGDP data, we kept the 5,000 SNPs with the largest Bayes factors. The first two factors mainly measure differentiation between Africa, Western Eurasia and East Asia and some outliers related with these two factors are involved in morphological traits, which is a trait often reported to be enriched with genes having signatures of positive selection (Barreiro et al. 2008). In the list of outliers, we found the genes *SLC24A5* and *EDAR*, often reported as top hits in selection scan (Pickrell et al. 2009; Hao et al. 2013), and related to skin pigmentation and hair thickness respectively (Kamberov et al. 2013; Mallick et al. 2013). We also found that SNPs close to the oncogene *CYP26A1* whose expression is enhanced in sunlight-damaged human skin (Osanai and Lee 2011) is part of the top list for outliers. The third and fourth factors correspond to genetic differentiation between Native Americans, individuals from Oceania and the rest of the sample. There are many regions in chromosome 8 enriched with outlier SNPs. One of this region encompasses the gene *RP1L1* that is associated with retinal diseases and which has already reported to have one of the strongest signature of positive selection along with other genes related to sensory functions (Barreiro et al. 2008). Many outlier SNPs strongly related to the third and fourth factors have allele frequencies that are similar between Southern Native Americans and Pygmies (third factor) or between individuals from Oceania, Southern Native Americans and Pygmies (fourth factor). Because these individuals all live in tropical rain forests and have similar diet consisting of roots and tubers, our findings support the importance of diet, climate, and potentially pathogen load to explain human adaptation (Hancock et al. 2010; Fumagalli et al. 2011). The SNPs with similar allele frequencies in different geographic regions are good candidates for convergent evolution and would deserve further analysis.

Factor models are enriching the toolbox of population genetic methods. The main principle is to model population structure via latent variables called factors. Factors models have already been proposed to ascertain population structure (Engelhardt and Stephens 2010), and to account for population structure when testing for gene-environment association (Frichot et al. 2013). We showed that factor models also provide a convenient individual-based framework to find loci that have atypical patterns of genetic differentiation. A major argument supporting the proposed hierarchical factor model is that it produces less false discoveries compared to genome scans based on F_{ST} .

Materials and Methods

Hierarchical Bayesian modeling

We provide the prior distributions for the latent variables of the hierarchical factor model defined by equations (1)-(3). To account for Linkage Disequilibrium (LD) in the genome and encourage outlier loci to be clustered along the genome, we consider a Potts model with external field for the indicator variables (Winkler 2003)

$$p(z_1, \dots, z_p) \propto (1 - \pi)^{p_0} \pi^{(p-p_0)} e^{\beta \sum_{i \sim j} \mathbb{1}_{z_i = z_j}}, \quad (4)$$

where the sum in the exponential ranges over all pairs of neighboring loci. We consider that each locus has two neighbors except at the beginning and at the end of the chromosome where a locus has only one neighbor. In equation (4), the variable p_0 is the number of loci such that $z_i = 0$, $\mathbb{1}$ is the indicator function, β is the parameter of the Potts model and is set to $\beta = 1$, and π is the prior proportion of outlier. To model the proportion of outlier, we consider a uniform prior on the \log_{10} scale reflecting that we are interested in the order of magnitude of the proportion of outlier loci (Guan and Stephens 2011). In the following we consider -4 and -1 for the lower and upper bound of the uniform prior. For the variance parameters σ_k^2 , $k = 1, \dots, K$, that are specific to each factor (equations (2) and (3)), we consider the parametrization $\sigma_k^2 = \sigma^2 \rho_k^2$ where σ^2 is the residual variance in equation (1) (Oba et al. 2003). We consider the non-informative prior for the variance parameters $p(\sigma^2) \propto 1/\sigma^2$ and $p(\rho_k^2) \propto 1/\rho_k^2$, $k = 1, \dots, K$. For the variance-inflation parameters c_k^2 of equation (3), we consider uniform priors with 1 and 10 for the lower and upper bounds.

With the factor model of equation (1), there is a well known issue of identifiability because identical likelihood values can be obtained from a solution (\mathbf{U}, \mathbf{V}) after using orthogonal rotations (West 2003). To add constraints to the model, we consider a prior with unit variance for each of the factors

$$\mathbf{U}_k \sim \mathcal{N}(0, \mathbf{I}_n),$$

where \mathbf{I}_n is the squared $n \times n$ identity matrix (Oba et al. 2003). To further prevent the MCMC algorithm to produce alternative rotations of the factors (Engelhardt and Stephens 2010), we consider the solution of the singular value decomposition as starting values for the factors $\mathbf{U}_1, \dots, \mathbf{U}_K$ in the MCMC algorithm.

To evaluate the strength of evidence for outlyingness at each locus, we compute the Bayes factor on a \log_{10} scale. The Bayes factor is defined as the ratio between the posterior odds $p(z_\ell > 0|Y)/p(z_\ell = 0|Y)$ and the prior odds $p(z_\ell > 0)/p(z_\ell = 0) = \pi/(1 - \pi)$. The description of the MCMC algorithm and of the computation of the Bayes factor is given in the supplementary material.

Simulation of the four-population divergence model

The first simulation scenario is a divergence model with 4 populations. Populations have a constant effective population sizes of $N_e = 1,000$ diploid individuals, with 50 individuals sampled in each population. The genotypes consist of 10,000 independent SNPs. The simulations are performed in two steps. In the first step, we use the software *ms* to simulate a neutral divergence model (Hudson 2002). When looking backward in time, we instantly merge population A_1 with A_2 , and population B_1 with B_2 , then after waiting a number $T = 20, 80, 120, 160, 200$ of generations, we merge the two remaining populations A and B. We keep only variants with a minor allele frequency larger than 5% at the end of this first step. The second step is performed with the software *SimuPOP* (Peng and Kimmel 2005). To run *SimuPOP*, we provide the allele frequencies in each of the 2 populations that have been generated with *ms*. Looking forward-in-time, we simulate 100 generations after the 2 concomitant divergence events. We assume no migration between populations. In each evolutionary lineage, we assume that 100 SNPs confer selective advantages using a selection coefficient of $s = .1$ for homozygotes carrying two adaptive alleles. In both simulation schemes, we assume an additive model for selection.

Simulation of the stepping-stone model

The second simulation scenario is a 2-dimensional stepping-stone model with a 10×10 grid. Each of the 100 populations has an effective population size of $N_e = 1,000$ diploid individuals. We sample 10 individuals in each population, and there are 2,050 independent SNPs. We also consider a 2-step procedure for the simulations. First, we simulate an equilibrium stepping-stone model with the software *ms*. Neighboring populations exchange migrants with a rate $4N_em = 8$ per generation. Then we superimpose a selection gradient using *SimuPOP*. During 100 generations, we consider that 50 SNPs confer selective advantage. The selection coefficient $s = 0.1$ is maximal in population 64, which is located in the lower right quarter of the grid. In the four neighboring population, the selection coefficient is of $s = 0.05$ and in the second layer of neighbors the selection coefficient is of $s = 0.025$. The selection coefficient is equal to 0 for the rest of the grid.

Gene ontology analysis

We perform a Gene Ontology (GO) enrichment analysis with *Gowinda* (Kofler and Schlötterer 2012). The list of genes is built by considering all genes which contain outlier SNPs with a tolerance of 5,000 base pairs upstream and downstream. We use a threshold of 0.05 for the false discovery rate, and we remove gene ontology terms that are shared by less than 10 genes or more than 1,000 genes. We consider the *-snp* flag in *Gowinda* that assumes independence of SNP within gene because the *-gene* flag, which assumes complete dependence of SNPs within a gene, does not provide any discovery with a FDR smaller than 5%. For each factor, we consider the ten gene ontology terms with the smallest

false discovery rates, and we only report gene ontology terms that are related to biological processes.

Software availability

The computer program *PCAdapt* for fitting the factor model is available from the authors' web-sites (<http://membres-timc.imag.fr/Michael.Blum/>, <http://membres-timc.imag.fr/Nicolas.Duforet-Frebourg/>).

Acknowledgments

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01).

Literature Cited

- Abdellaoui, A., J.-J. Hottenga, P. de Knijff, M. G. Nivard, X. Xiao, P. Scheet, A. Brooks, E. A. Ehli, Y. Hu, G. E. Davies, et al., 2013. Population structure, migration, and diversifying selection in the Netherlands. *European Journal of Human Genetics* 21:1277–1285.
- Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome research* 12:1805–1814.
- Antao, T., A. Lopes, R. Lopes, A. Beja-Pereira, and G. Luikart, 2008. Lositan: a workbench to detect molecular adaptation based on a F_{ST} -outlier method. *BMC bioinformatics* 9:323.
- Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, 2008. Natural selection has driven population differentiation in modern humans. *Nature genetics* 40:340–345.
- Bazin, E., K. J. Dawson, and M. A. Beaumont, 2010. Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* 185:587–602.
- Beaumont, M. A. and D. J. Balding, 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology* 13:969–980.
- Beaumont, M. A. and R. A. Nichols, 1996. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 263:1619–1626.
- Bierne, N., D. Roze, and J. J. Welch, 2013. Pervasive selection or is it...? why are F_{ST} outliers sometimes so frequent? *Molecular ecology* 22:2061–2064.
- Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. SanCristobal, 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186:241–262.

- Bonin, A., P. Taberlet, C. Miaud, and F. Pompanon, 2006. Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution* 23:773–783.
- Box, G. E. and G. C. Tiao, 1968. A Bayesian approach to some outlier problems. *Biometrika* 55:119–129.
- Canfield, V. A., A. Berg, S. Peckins, S. M. Wentzel, K. C. Ang, S. Oppenheimer, and K. C. Cheng, 2013. Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3: Genes| Genomes| Genetics* 3:2059–2067.
- Davidson, A. E., P. I. Sergouniotis, D. S. Mackay, G. A. Wright, N. H. Waseem, M. Michaelides, G. E. Holder, A. G. Robson, A. T. Moore, V. Plagnol, et al., 2013. RP1L1 variants are associated with a spectrum of inherited retinal diseases including retinitis pigmentosa and occult macular dystrophy. *Human mutation* 34:506–514.
- Devlin, B. and K. Roeder, 1999. Genomic control for association studies. *Biometrics* 55:997–1004.
- Engelhardt, B. E. and M. Stephens, 2010. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS genetics* 6:e1001117.
- Evanno, G., S. Regnaut, and J. Goudet, 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology* 14:2611–2620.
- Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin, 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193:929–941.
- Foll, M. and O. Gaggiotti, 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180:977–993.
- Fourcade, Y., A. Chaput-Bardy, J. Secondi, C. Fleurant, and C. Lemaire, 2013. Is local selection so widespread in river organisms? fractal geometry of river networks leads to high bias in outlier detection. *Molecular ecology* .
- Frichot, E., S. D. Schoville, G. Bouchard, and O. François, 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular biology and evolution* 30:1687–1699.
- Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admettla, L. Pattini, and R. Nielsen, 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS genetics* 7:e1002355.
- Gompert, Z. and C. A. Buerkle, 2011. A hierarchical Bayesian model for next-generation population genomics. *Genetics* 187:903–917.

- Guan, Y. and M. Stephens, 2011. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* 5:1780–1815.
- Günther, T. and G. Coop, 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195:205–220.
- Guo, F., D. K. Dey, and K. E. Holsinger, 2009. A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association* 104:142–154.
- Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, R. Sukernik, G. Utermann, J. Pritchard, G. Coop, et al., 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences* 107:8924–8930.
- Hao, W., M. Song, and J. D. Storey, 2013. Probabilistic models of genetic variation in structured populations applied to global human studies. *arXiv preprint arXiv:1312.2041*.
- Hilborn, R. and M. Mangel, 1997. *The ecological detective: confronting models with data*, vol. 28. Princeton University Press.
- Hudson, R. R., 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
- Jarvis, J. P., L. B. Scheinfeldt, S. Soi, C. Lambert, L. Omberg, B. Ferwerda, A. Froment, J.-M. Bodo, W. Beggs, G. Hoffman, et al., 2012. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics* 8:e1002641.
- Kamberov, Y. G., S. Wang, J. Tan, P. Gerbault, A. Wark, L. Tan, Y. Yang, S. Li, K. Tang, H. Chen, et al., 2013. Modeling recent human evolution in mice by expression of a selected *edar* variant. *Cell* 152:691–702.
- Kofler, R. and C. Schlötterer, 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* 28:2084–2085.
- Laloë, D. and M. Gautier, 2011. On the genetic interpretation of between-group PCA on SNP data. HAL hal-00661214.
- Lange, K., J. C. Papp, J. S. Sinsheimer, and E. M. Sobel, 2014. Next-generation statistical genetics: Modeling, penalization, and optimization in high-dimensional data. *Annual Review of Statistics and Its Application* 1:279–300.
- Lewontin, R. and J. Krakauer, 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175–195.

- Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, et al., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
- Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet, 2003. The power and promise of population genomics: from genotyping to genome typing. *Nature reviews genetics* 4:981–994.
- Mallick, C. B., F. M. Iliescu, M. Möls, S. Hill, R. Tamang, G. Chaubey, R. Goto, S. Y. Ho, I. G. Romero, F. Crivellaro, et al., 2013. The light skin allele of *SLC24A5* in South Asians and Europeans shares identity by descent. *PLoS genetics* 9:e1003912.
- Manel, S. and R. Holderegger, 2013. Ten years of landscape genetics. *Trends in Ecology & Evolution* 28:614–621.
- Nicholson, G., A. V. Smith, F. Jónsson, Ó. Gústafsson, K. Stefánsson, and P. Donnelly, 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64:695–715.
- Novembre, J. and M. Stephens, 2008. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* 40:646–649.
- Oba, S., M. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, 2003. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19:2088–2096.
- Oleksyk, T. K., M. W. Smith, and S. J. O’Brien, 2010. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365:185–205.
- Osanai, M. and G.-H. Lee, 2011. Enhanced expression of retinoic acid-metabolizing enzyme *cyp26a1* in sunlight-damaged human skin. *Medical molecular morphology* 44:200–206.
- Patterson, N., A. L. Price, and D. Reich, 2006. Population structure and eigenanalysis. *PLoS Genetics* 2:e190.
- Peng, B. and M. Kimmel, 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21:3686–3687.
- Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivasan, G. S. Barsh, R. M. Myers, M. W. Feldman, et al., 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome research* 19:826–837.
- Rajeevan, H., U. Soundararajan, J. R. Kidd, A. J. Pakstis, and K. K. Kidd, 2012. ALFRED: an allele frequency resource for research and teaching. *Nucleic acids research* 40:D1010–D1015.

- Riebler, A., L. Held, and W. Stephan, 2008. Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178:1817–1829.
- Vitalis, R., K. Dawson, and P. Boursot, 2001. Interpretation of variation across marker loci as evidence of selection. *Genetics* 158:1811–1823.
- Vitalis, R., K. Dawson, P. Boursot, and K. Belkhir, 2003. DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity* 94:429–431.
- Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al., 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42:D1001–D1006.
- West, M., 2003. Bayesian factor regression models in the “large p, small n” paradigm. *Bayesian statistics* 7:723–732.
- Winkler, G., 2003. Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction, vol. 27. Springer.
- Wright, S., 1931. Evolution in Mendelian populations. *Genetics* 16:97.
- Xu, S., S. Li, Y. Yang, J. Tan, H. Lou, W. Jin, L. Yang, X. Pan, J. Wang, Y. Shen, et al., 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Molecular biology and evolution* 28:1003–1011.
- Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, et al., 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329:75–78.

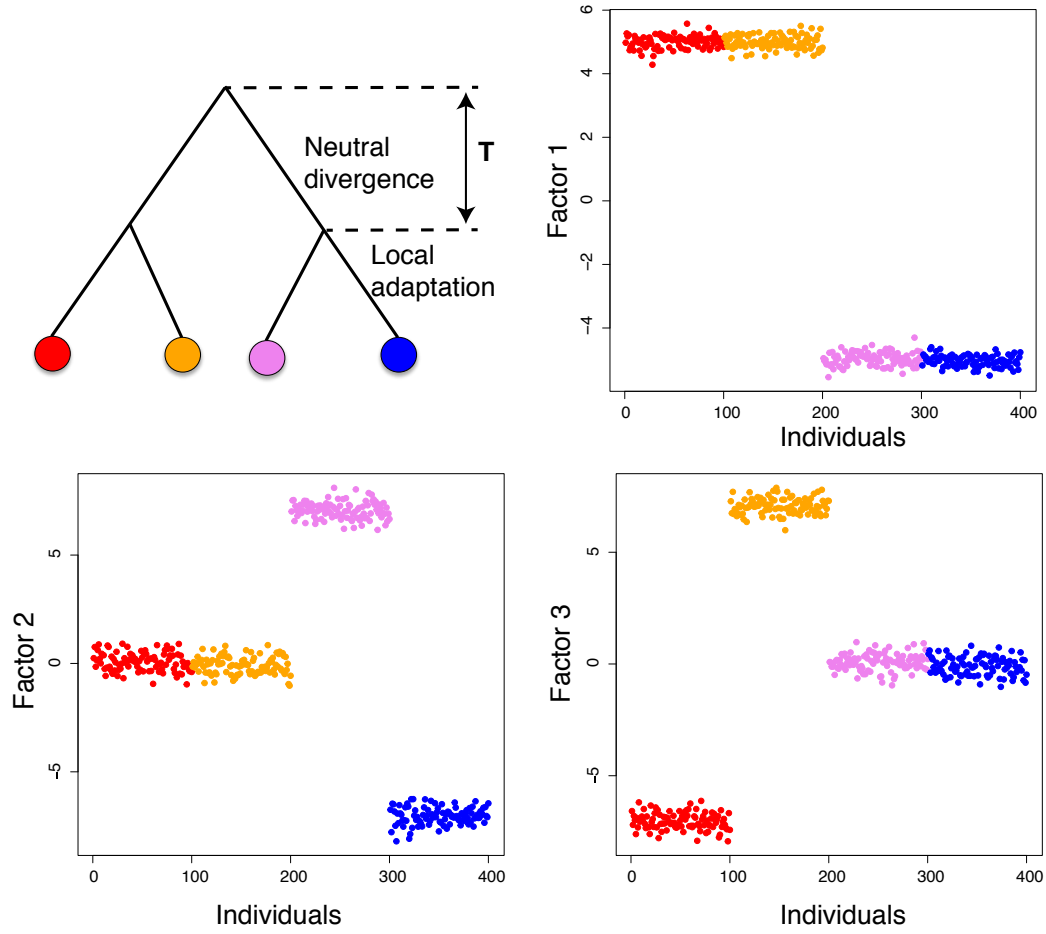


Figure 1: Values of the $K = 3$ factors for a population divergence model with 4 populations. The upper left panel shows the models of population divergence. The other panels show the values of the first three factors and each dot corresponds to one individual. As candidates for local adaptation, the factor model with $K = 3$ looks for SNPs whose variation is atypically well explained by one of the three factors.

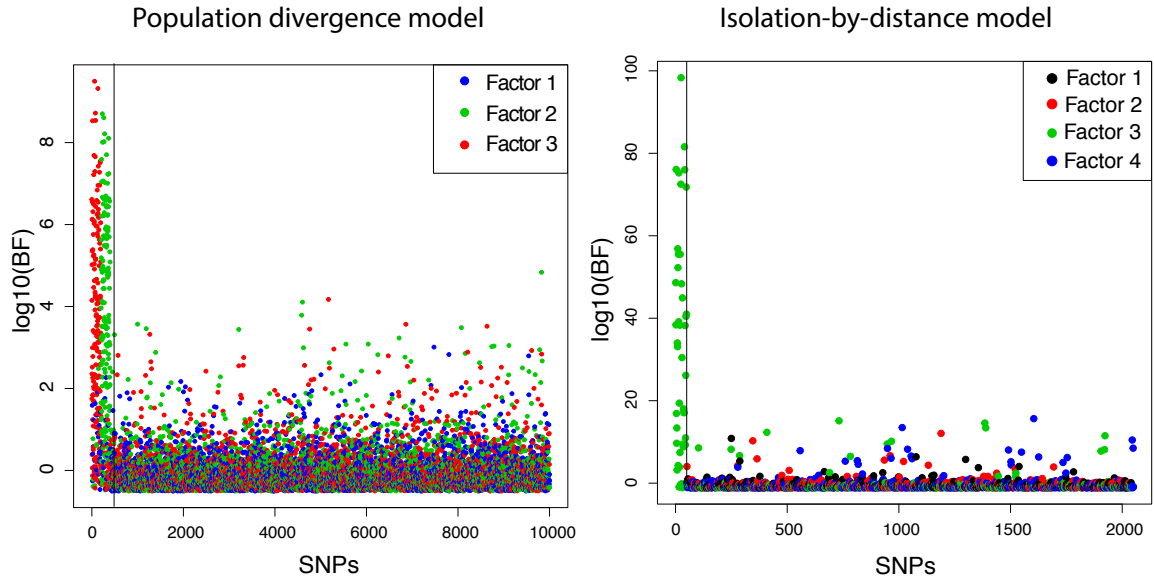


Figure 2: Bayes factors for the two different examples. The SNPs under selection are located on the left-hand side of the vertical bar.

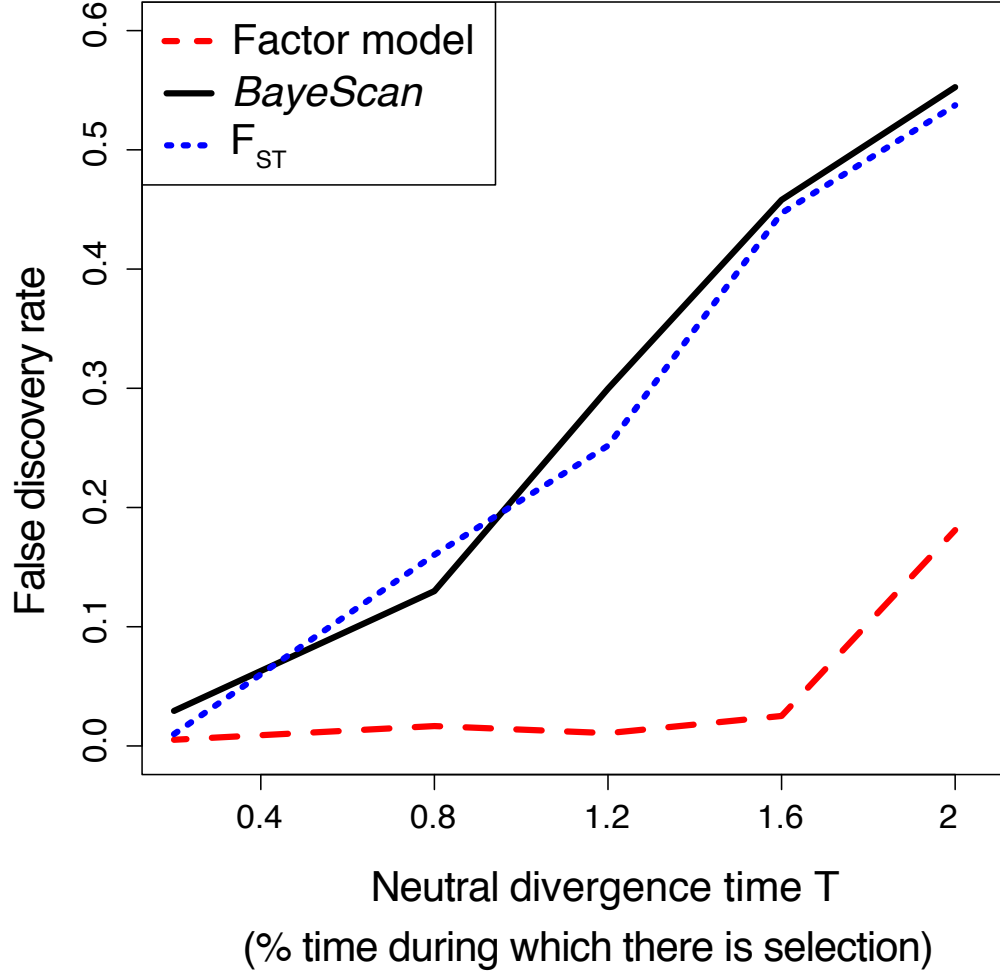


Figure 3: False discovery rate as a function of the initial divergence time T in the population divergence model of Figure 1. For both *BayeScan* and the proposed factor model, Bayes factors are used for ranking SNPs whereas we use F_{ST} values for the standard genome scan based on F_{ST} values. To determine a threshold above which SNPs are considered as outlier, we constrain the lists of SNPs provided by each method to contain 50% of the 400 SNPs truly involved in local adaptation. The neutral divergence time T is scaled so that $T = 1$ means that the neutral and adaptive phases are of same duration.

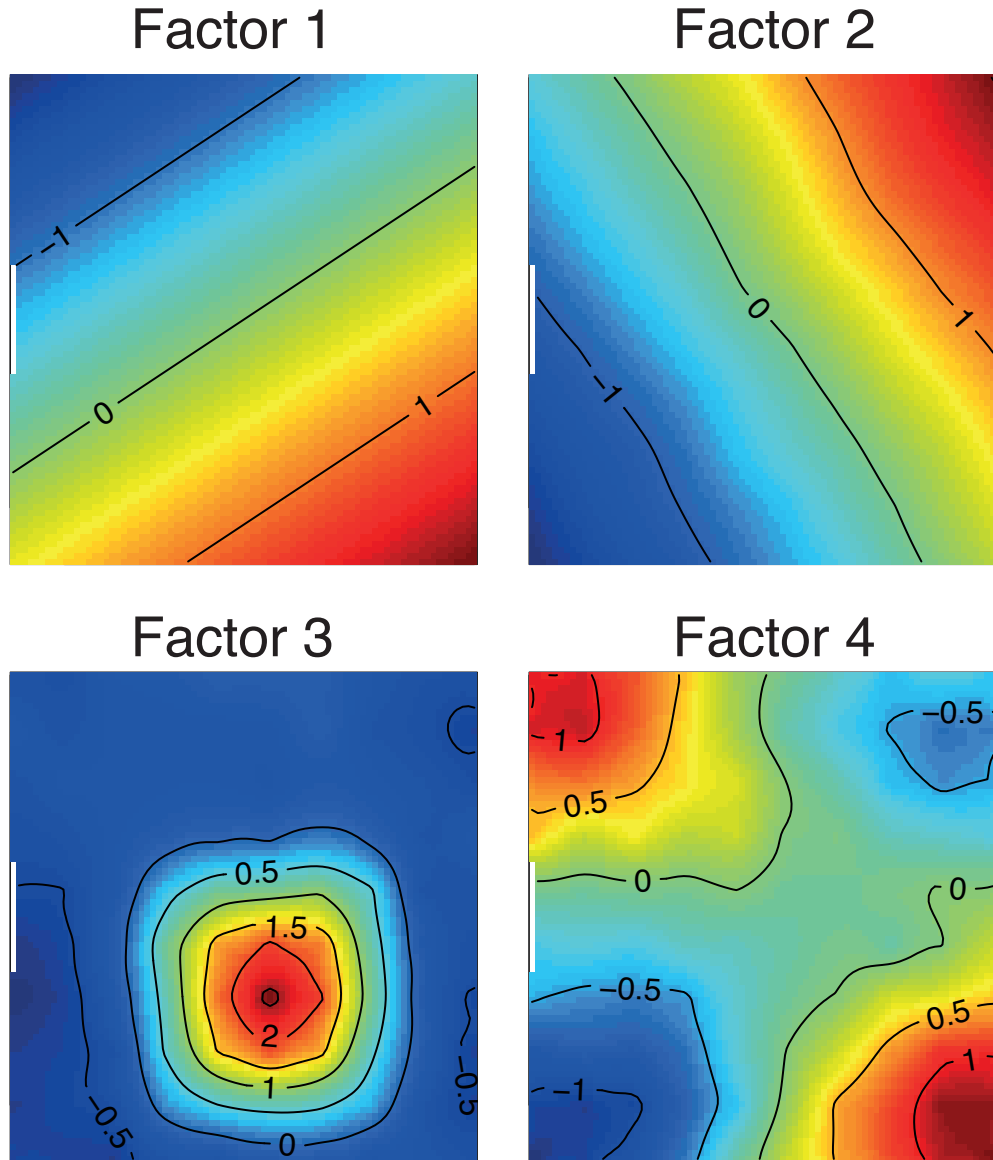


Figure 4: Values of the $K = 4$ factors for the two dimensional isolation-by-distance model. The different panels shows the values of the first four factors when projected onto the two-dimensional space. As candidates for local adaptation, the factor model with $K = 4$ looks for SNPs whose variation is atypically well explained by one of the four factors. Spatial interpolation of the factors is obtained using the *Krig* function that is available from the *fields* R package.

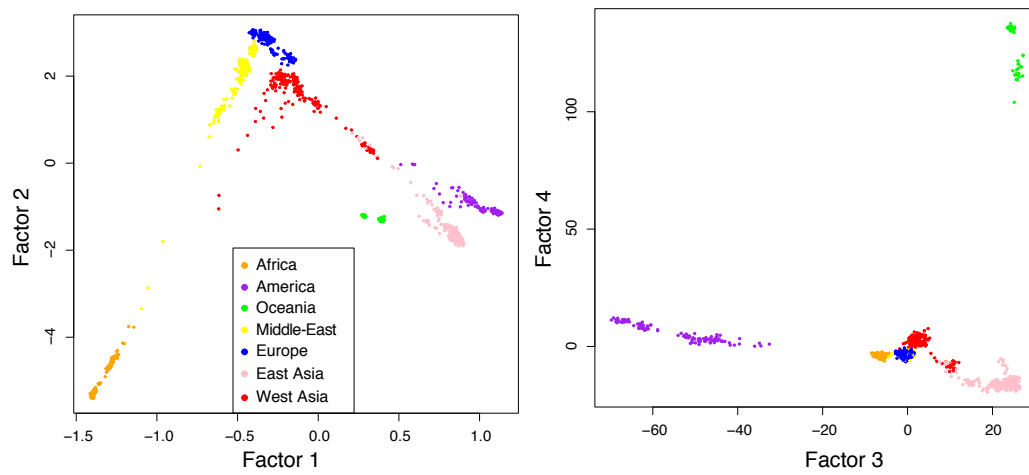


Figure 5: Values of the $K = 4$ factors for the HGDP dataset.

Chromosome	rs identifier	closest gene	dist.*	$\log_{10}(\text{BF})^+$	Factor no.
chr 10	4918664	<i>CYP26A1</i>	83	2.8	1
chr 10	10882168	<i>CYP26A1</i>	92	2.7	1
chr 10	7091054	<i>MYOF</i>	48	2.7	1
chr 10	11187300	<i>CYP26A1</i>	83	2.6	1
chr 2	7556886	<i>SMC6</i>	0	2.6	1
chr 10	12220128	<i>MYOF</i>	91	2.5	1
chr 10	6583859	<i>CYP26A1</i>	56	2.5	1
chr 10	4918924	<i>MYOF</i>	89	2.4	1
chr 2	1834619	<i>SMC6</i>	0	2.4	1
chr 2	4578856	<i>SMC6</i>	0	2.4	1
chr 15	1834640	<i>SLC24A5</i>	21	5.1	2
chr 15	2250072	<i>SLC24A5</i>	29	4.2	2
chr 2	260714	<i>EDAR</i>	0	3.2	2
chr 1	7531501	<i>SLC35F3</i>	0	2.9	2
chr 15	11637235	<i>DUT</i>	0	2.9	2
chr 9	10760260	<i>RABGAP1</i>	0	2.9	2
chr 9	2416899	<i>STRBP</i>	0	2.8	2
chr 5	2406410	<i>KIF3A</i>	0	2.8	2
chr 15	3751631	<i>MYO5C</i>	0	2.8	2
chr 9	618746	<i>RABGAP1</i>	0	2.8	2
chr 22	139553	<i>MEI1</i>	0	4.1	3
chr 22	5996039	<i>PMM1</i>	0	4.0	3
chr 22	8139993	<i>DESI1</i>	0	4.0	3
chr 22	126092	<i>MEI1</i>	0	4.0	3
chr 22	1005402	<i>XPNPEP3</i>	0	3.6	3
chr 22	8137373	<i>ZC3H7B</i>	0	3.6	3
chr 22	133074	<i>MCHR1</i>	0	3.6	3
chr 22	9611613	<i>CSDC2</i>	0	3.5	3
chr 20	2424641	<i>SYNDIG1</i>	71	3.5	3
chr 14	2600814	<i>LINC00871</i>	13	3.5	3
chr 8	16892216	<i>MAL2</i>	19	2.8	4
chr 8	6990312	<i>SYBU</i>	0	2.8	4
chr 17	9908046	<i>MMD</i>	64	2.7	4
chr 17	575873	<i>MEOX1</i>	18	2.6	4
chr 4	4691075	<i>NPY1R</i>	0	2.6	4
chr 17	4471745	<i>MMD</i>	70	2.5	4
chr 14	12891534	<i>CEP128</i>	0	2.5	4
chr 8	6988341	<i>SYBU</i>	2	2.5	4
chr 8	12216712	<i>MSRA</i>	16	2.5	4
chr 17	11869714	<i>MYCBPAP</i>	0	2.5	4

* dist. is the distance from the closest gene and is measured in kilo base pairs.

⁺ $\log_{10}(\text{BF})$ is the logarithm (in base 10) of the Bayes factor

Table 1: List of the 10 SNPs with largest Bayes factors for each of the four factors obtained with the HGDP dataset.

Supplementary Material of "Genome scans for detecting footprints of local adaptation using a Bayesian factor model."

- Description of the MCMC algorithm and of the computation of the Bayes factors.
- Figure S1: Population divergence model: false discovery rate as a function of the initial divergence time when the sensitivity (recall) has been set to 25% and 75%.
- Figure S2: Population divergence model: false discovery rate obtained with the factor model using different values of the number of factors K .
- Figure S3: Two example models and HGDP: average mean square error as a function of the number of factors K .
- Figure S4: IBD model: false discovery rate as a function of the number of factors K .
- Figure S5: IBD model: false discovery rate as a function of the number of sampled individuals per deme.
- Figure S6: HGDP dataset: Plot of factors 5-8.
- Figure S7: HGDP dataset: Number of outlier SNPs per chromosome.
- Table S1: HGDP dataset: List of the 5,000 SNPs with largest Bayes factors (csv file).
- Table S2: HGDP dataset: List of the 4×10 SNPs with largest Bayes factors and links to the *ALFRED* database to display allele frequencies.
- Table S3: Results of the Gene Ontology based on the set of genes found in Table S1.
- Table S4: List of GWAS SNPs that are found for the outlier SNPs (csv file).

MCMC algorithm

We model the variance parameter of the regression coefficients using the parametrization $\sigma_k^2 = \sigma^2 \rho_k^2$. In the following, we consider $\Sigma_V = \sigma^2 \times \text{Diag}(\rho_1^2, \dots, \rho_K^2)$ where $\text{Diag}(\rho_1^2, \dots, \rho_K^2)$ is a diagonal matrix with $(\rho_1^2, \dots, \rho_K^2)$ in the diagonal. Σ_{V,k_0} is the same as Σ_V except that the k_0^{th} element in the diagonal is $\sigma^2 c_{k_0}^2 \rho_{k_0}^2$ instead of $\sigma^2 \rho_{k_0}^2$. For a given value of K , we describe below a single step of the MCMC algorithm where we use Gibbs updating steps except when updating c_1^2, \dots, c_K^2 and π .

-

$$\mathbf{V}_\ell \leftarrow \mathcal{N}(\mathbf{m}_\ell, \Sigma_\ell), \ell = 1, \dots, p,$$

$$\text{with } \mathbf{m}_\ell = (\sigma^2 \Sigma_{V,k_0}^{-1} + \mathbf{U}^T \mathbf{U})^{-1} (\mathbf{U} \mathbf{Y}_\ell) \text{ and } \Sigma_\ell = (\Sigma_{V,k_0}^{-1} + \frac{1}{\sigma^2} \mathbf{U}^T \mathbf{U})^{-1}.$$

-

$$\rho_k^2 \leftarrow IG\left(\frac{p}{2}, \frac{1}{2\sigma^2} \sum_{\ell=1}^p \frac{V_{k\ell}^2}{(c_k^2)^{\mathbb{1}_{z_\ell=k}}}\right),$$

where $\mathbb{1}$ is the indicator function and IG is the inverse-gamma distribution.

-

$$\sigma^2 \leftarrow IG\left(\frac{(n+K)p}{2}, \frac{1}{2} \sum_{\ell=1}^p \left(\sum_{i=1}^n |Y_{i,\ell} - U_i V_\ell|^2 + \sum_{k=1}^K \frac{V_{k\ell}^2}{\rho_k^2 (c_k^2)^{\mathbb{1}_{z_\ell=k}}} \right)\right),$$

-

$$\mathbf{U}_i \leftarrow \mathcal{N}(\mathbf{m}_i, \Sigma), i = 1, \dots, n,$$

with $\mathbf{m}_i = (I_K + \frac{1}{\sigma^2} \mathbf{V} \mathbf{V}^T)^{-1} \frac{1}{\sigma^2} \mathbf{V} \mathbf{Y}_i$ and $\Sigma = (\frac{1}{\sigma^2} \mathbf{V} \mathbf{V}^T + I_K)^{-1}$. The \mathbf{U}_i 's are vectors of dimension K and are the row vectors of the matrix \mathbf{U} . Note that in the main text, we rather refer to the column vectors of the matrix \mathbf{U} when using the notation \mathbf{U}_i .

-

$$z_\ell = k \text{ with a probability } p(z_\ell = k | \pi, V_\ell, \Sigma_V, c_k^2), k = 1, \dots, K$$

with

$$p(z_\ell = k | \pi, V_\ell, \Sigma_V, c_k^2) \propto \frac{\pi}{K} e^{\beta \sum_{i \sim \ell} \mathbb{1}_{z_i, z_\ell}} \frac{1}{c} e^{\frac{V_{k\ell}^2}{2\sigma^2 \rho_k^2} (1 - 1/c_k^2)}, \quad (5)$$

and $p(z_\ell = 0 | \pi, V_\ell, \Sigma_V, c_k^2) \propto (1 - \pi) e^{\beta \sum_{i \sim \ell} \mathbb{1}_{z_i, z_\ell}}$ for $k = 0$.

- Update (c_1^2, \dots, c_K^2) using a Metropolis-Hasting step where the proposal for each element is a Gaussian random walk of variance equal to 1.

- Update $\log_{10}(\pi)$ using a Metropolis-Hasting step where the proposal is a Gaussian random walk of variance equal to 0.5.

We find that MCMC runs of 400 iterations (burn-in of 200 iterations) are long-enough to provide convergence.

Bayes factors

For each SNP, the Bayes factor is defined as the ratio between the posterior odds (outlier vs non-outlier) and the prior odds $\pi/(1 - \pi)$. For numerical convenience, we rather compute the logarithm (in base 10) of the Bayes factor using a Monte-Carlo approximation. The \log_{10} of the posterior odds (LPO) can be computed along the MCMC run by averaging

$$LPO = \log_{10}\left\{\sum_{k=1}^K p(z_\ell = k|\cdot)\right\} - \log_{10}\{p(z_\ell = 0|\cdot)\}, \quad \ell = 1, \dots, p. \quad (6)$$

where $p(z_\ell|\cdot)$ is the conditional probability of z_ℓ given the other parameter values. Because the maximum of the probabilities $p(z_\ell = 1|\cdot), \dots, p(z_\ell = K|\cdot)$ is on a much larger scale compared to the other probabilities, we approximate equation (6) using

$$LPO \sim \log_{10}\left\{\max_{k=1, \dots, K} p(z_\ell = k|\cdot)\right\} - \log_{10}\{p(z_\ell = 0|\cdot)\},$$

and the conditional probabilities $p(z_\ell|\cdot)$ are given in the equation (5) of the supplementary material.

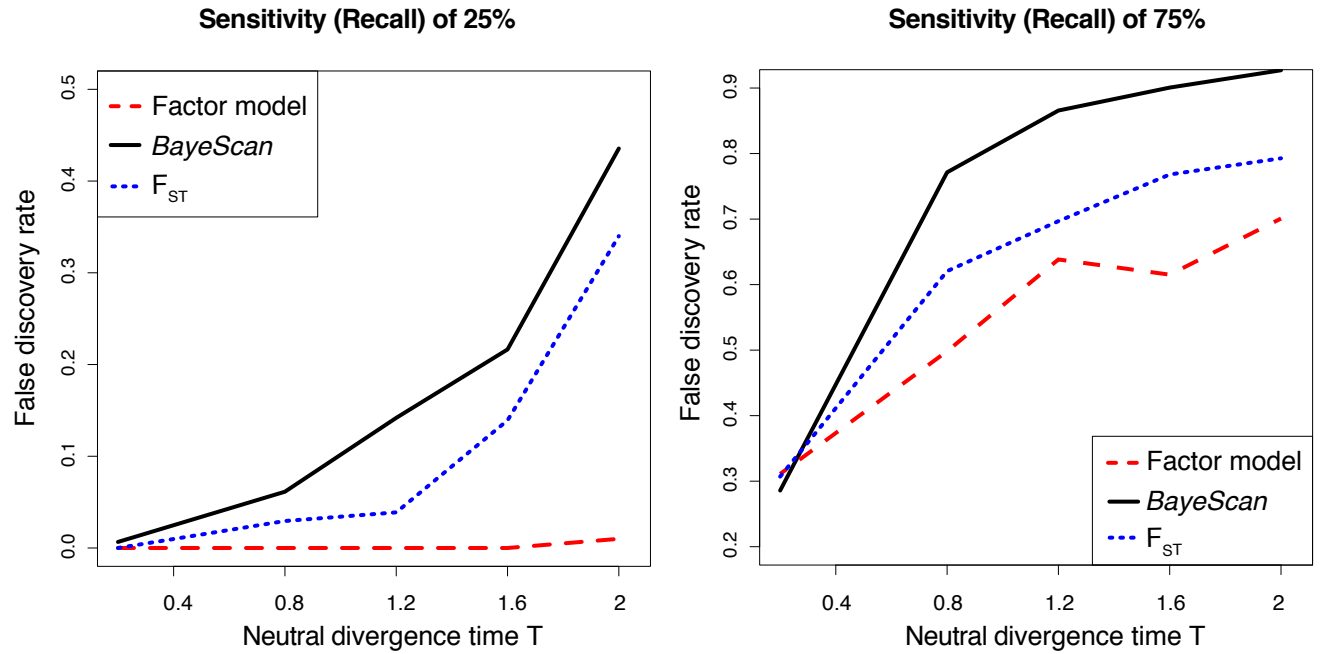


Figure S1: False discovery rate as a function of the initial divergence time in the population divergence model depicted in the Figure 1 of the main text. For both *BayeScan* and the proposed factor model, Bayes factors are used for ranking SNPs whereas we use F_{ST} values for genome scans based on F_{ST} values. To determine a threshold above which SNPs are considered as outlier, we constrain the lists of top-ranked SNPs provided by each method to contain 25%(left panel) or 75% (right panel) of the 400 SNPs truly involved in local adaptation. Time is counted in units of time during which there is selection.

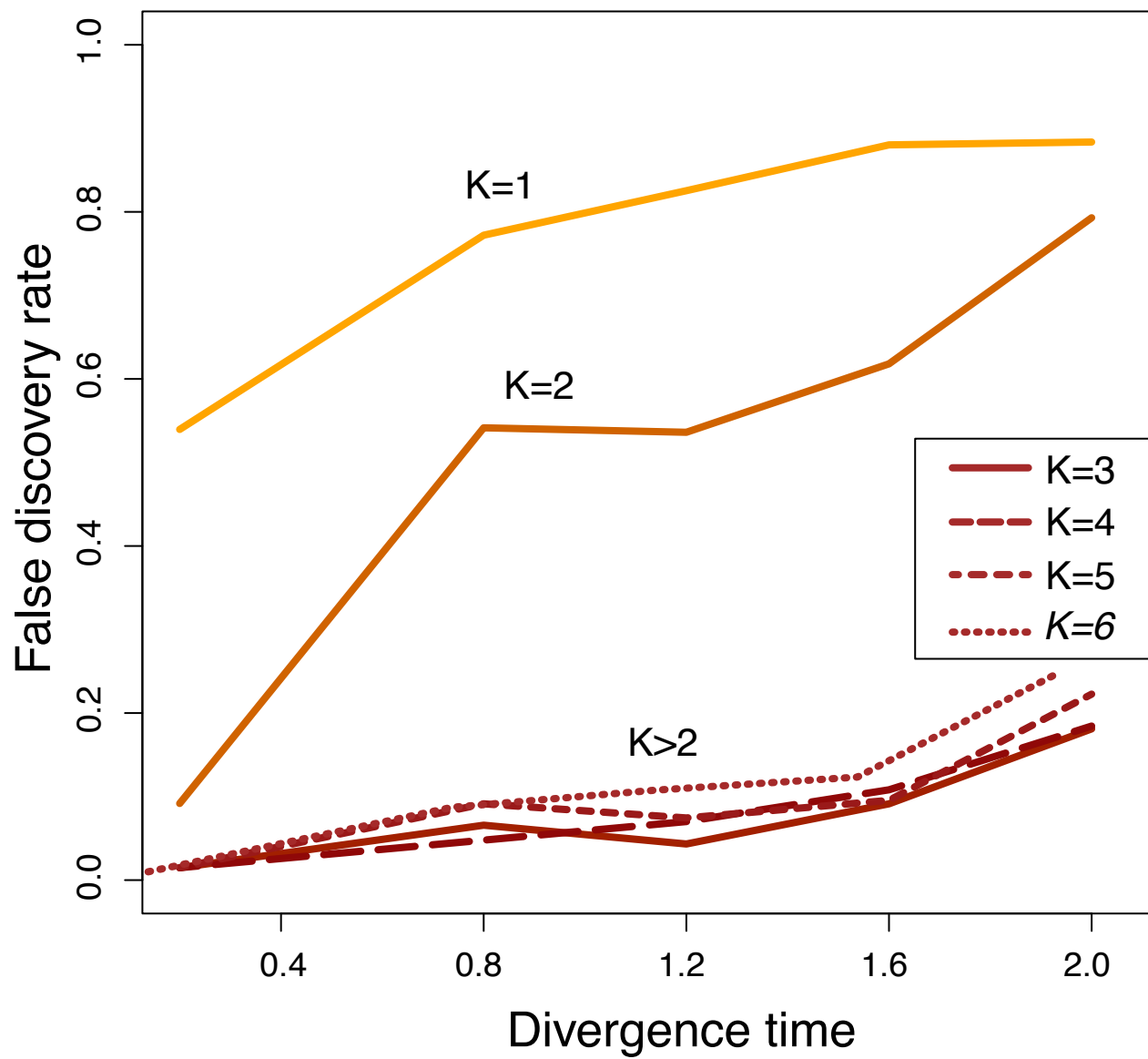


Figure S2: False discovery rate obtained with the factor model using different values of the number of factors K . Data are simulated with the population divergence model of the Figure 1 of the main text. To determine a threshold above which SNPs are considered as outlier, we constrain the lists of top-ranked SNPs provided by each method to contain 50% of the 400 SNPs truly involved in local adaptation.

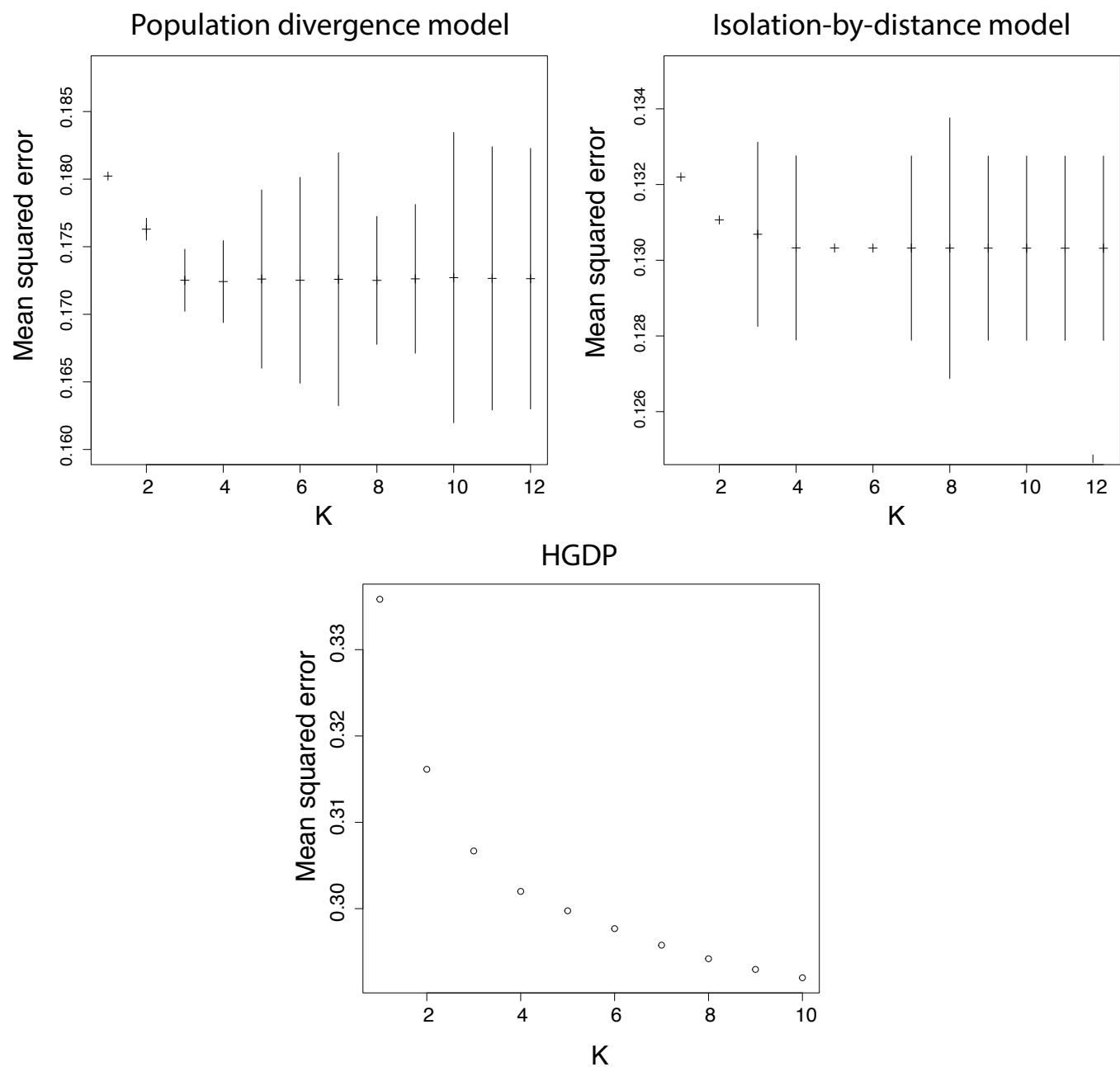


Figure S3: Average mean square error as a function of the number of factors K . Error bars correspond to 2 standard deviations and are estimated using a total of 10 different MCMC runs.

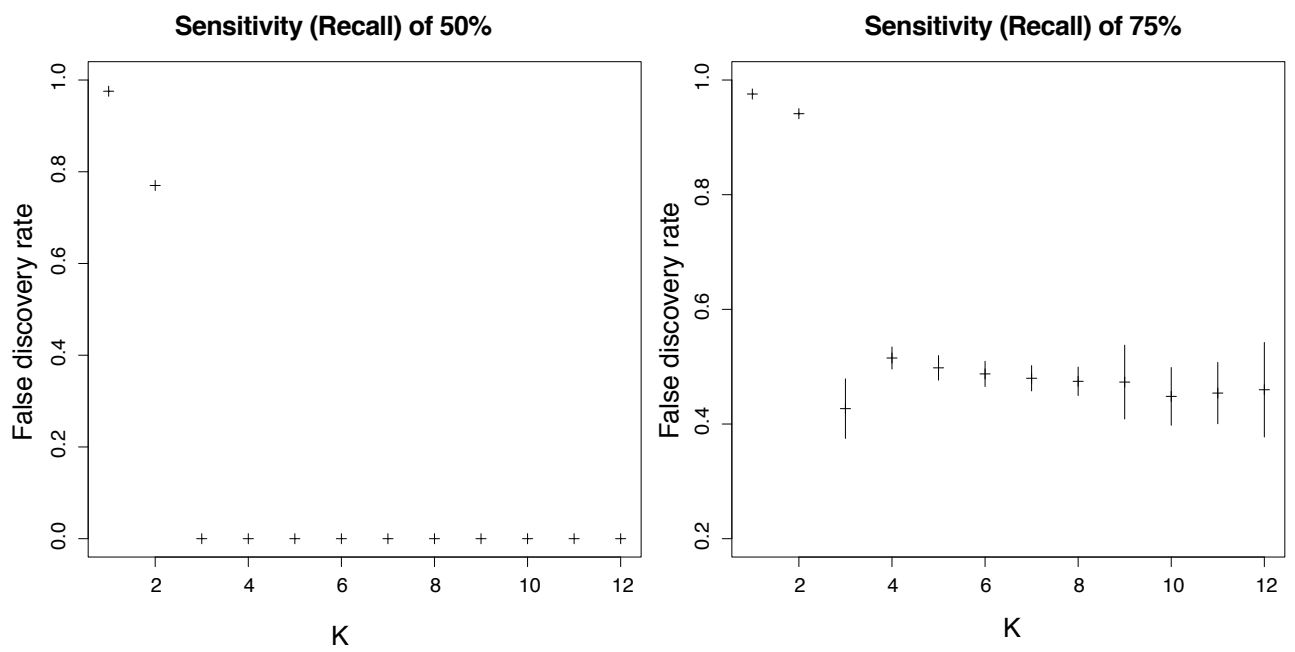


Figure S4: False discovery rate obtained with the factor model using different values of the number of factors K in the IBD model. Error bars correspond to 2 standard deviations and are estimated using a total of 10 different MCMC runs.

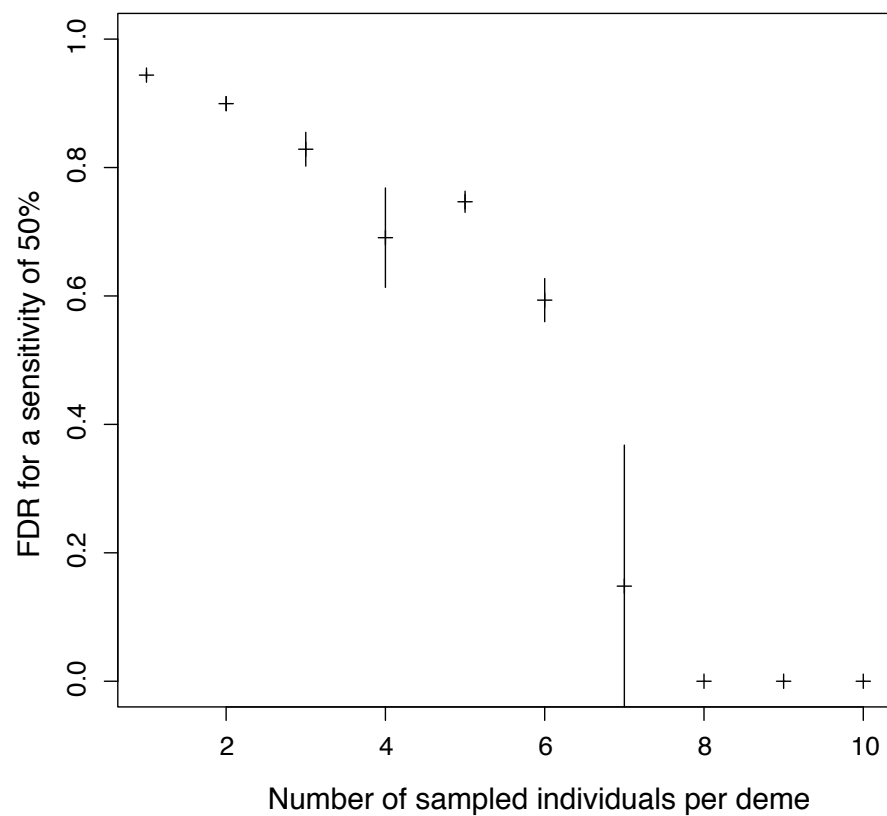


Figure S5: False discovery rate as a function of the number of sampled individuals per deme for the IBD model.

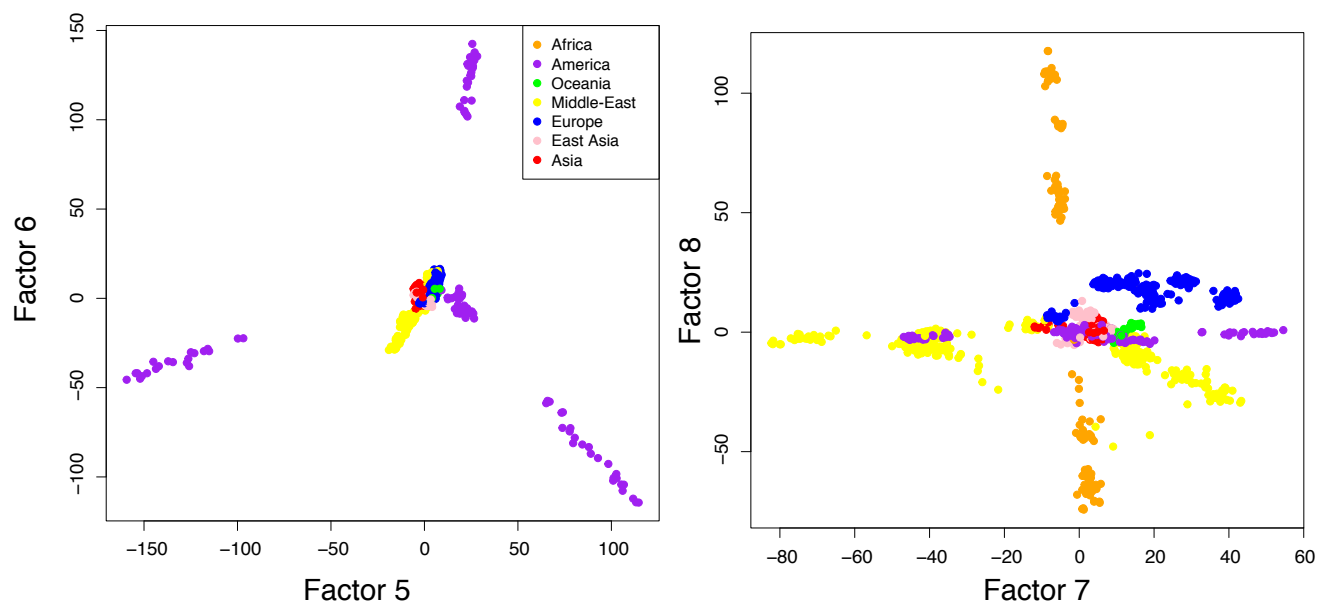


Figure S6: Values of factors 5 – 8 for the HGDP dataset.

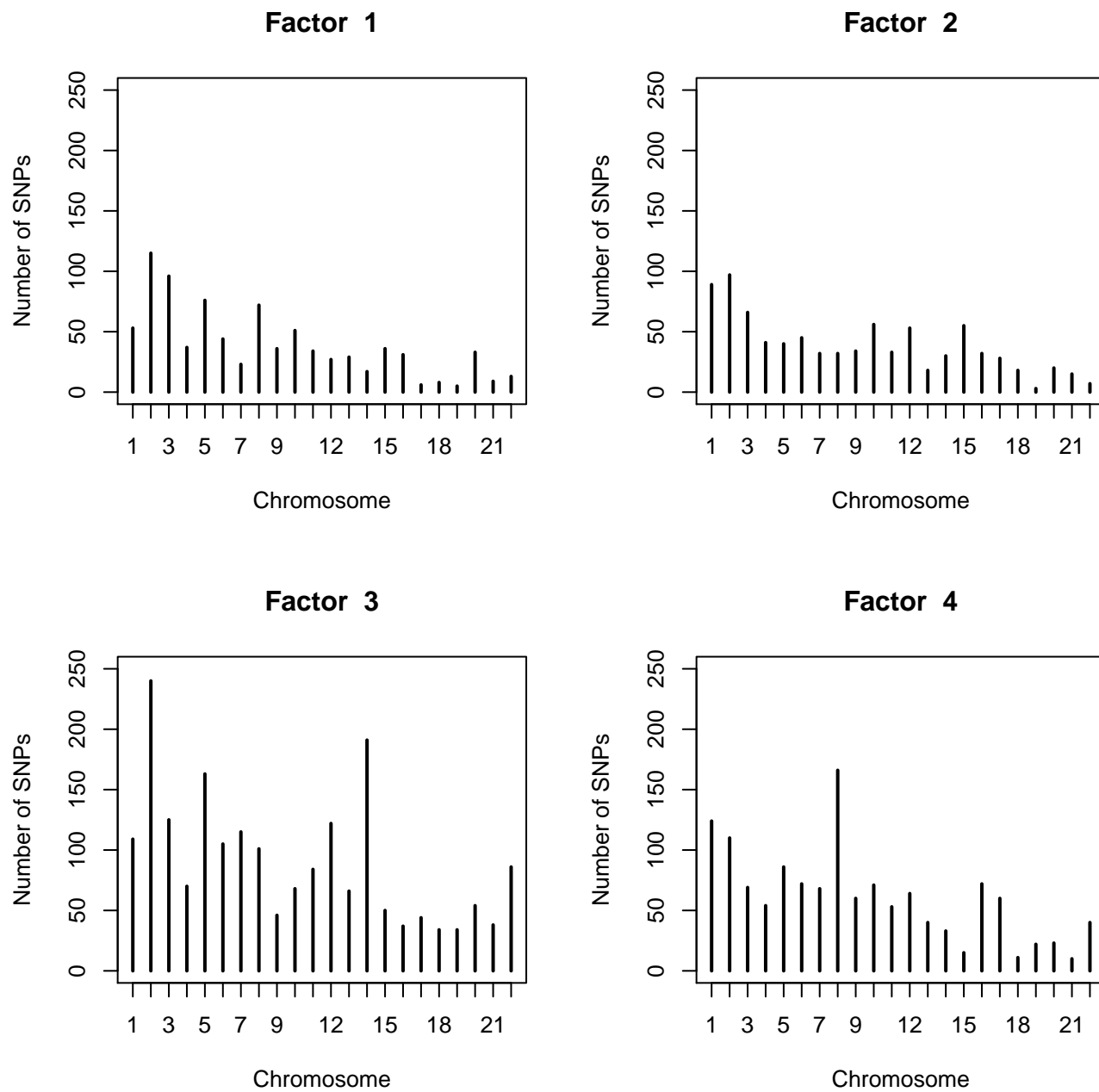


Figure S7: Number of outlier SNPs per chromosome for each of the four different factors.

rs identifier	Factor	Freq. Map
4918664	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94911055&chr=chr10&rs=rs4918664&imp=false
10882168	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94919424&chr=chr10&rs=rs10882168&imp=false
7091054	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI533583B
11187300	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI091645Z
7556886	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI563914C
12220128	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI136240Q
6583859	1	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=94883463&chr=chr10&rs=rs6583859&imp=false
4918924	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI450066V
1834619	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI274807C
4578856	1	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI421641S
1834640	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=46179457&chr=chr15&rs=rs1834640&imp=false
2250072	2	http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=46172199&chr=chr15&rs=rs2250072&imp=false
260714	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI015078V
7531501	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI561912Y
11637235	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI103044M
10760260	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI059055Y
2416899	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI333938D
2406410	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI333101L
3751631	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI386311W
618746	2	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI478518H
139553	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI197254C
5996039	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI469320Y
8139993	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI610594Z
126092	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI156639E
1005402	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI022498Z
8137373	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI610490U
133074	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI184162W
9611613	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI645506A
2424641	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI334632V
2600814	3	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI345176A
16892216	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI231611O
6990312	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI525364Z
9908046	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI659032Z
575873	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI014358V
4691075	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI428601V
4471745	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI416772B
12891534	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI168020R
6988341	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI525168B
12216712	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI136091U
11869714	4	http://alfred.med.yale.edu/alfred/mvograph.asp?siteuid=SI014860T

Table S2: List of the 10 SNPs with largest Bayes factors for each of the four factors obtained with the HGDP dataset. For each SNP, we provide the link to the *ALFRED* database to display maps or bar charts of allele frequencies.

Gene Ontology	no. SNP*	no. genes ⁺	on chip ⁻	GO description	Factor
GO:0001522	6	2	15	pseudouridine synthesis	1
GO:0030324	19	5	79	lung development	1
GO:0032024	9	3	28	positive regulation of insulin secretion	1
GO:0090277	9	3	35	positive regulation of peptide hormone secretion	1
GO:0046887	9	3	51	positive regulation of hormone secretion	1
GO:0008045	11	1	16	motor axon guidance	1
GO:0051896	9	4	56	regulation of protein kinase B signaling cascade	1
GO:0042993	6	1	18	positive regulation of transcription factor import into nucleus	2
GO:0042346	6	1	14	positive regulation of NF-kappaB import into nucleus	2
GO:0007156	61	12	115	homophilic cell adhesion	3
GO:0009896	30	4	97	positive regulation of catabolic process	3
GO:0031331	29	3	76	positive regulation of cellular catabolic process	3
GO:0000096	20	1	29	sulfur amino acid metabolic process	4
GO:0042982	10	1	13	amyloid precursor protein metabolic process	4
GO:0006516	10	1	15	glycoprotein catabolic process	4
GO:0006555	20	1	17	methionine metabolic process	4
GO:0007568	26	5	175	aging	4

* number of outlier SNPs

⁺ number of outlier genes

⁻ number of genes on SNP array

Table S3: Results of the Gene Ontology based on the set of genes found in Table S1. We use a threshold of 0.05 for the false discovery rate, we remove gene ontology terms that are shared by less than 10 genes or more than 1,000 genes. For each factor, we consider the ten GOs with the smallest FDRs and we only report the gene ontology terms that are related to biological processes.